Statistical Analysis in Information Assurance

A Presentation to

The Naval Post-Graduate School
Monterey, California

January 6, 2005

Daniel J. Ryan
ryand@ndu.edu

National Defense University

The Risk Management Equation

$$\text{Risk} = \frac{\text{Threats X Vulnerabilities}}{\text{Countermeasures}} \text{ X Impact}$$

Daniel J. Ryan
ryand@ndu.edu

National Defense University

Managing Risk

- **Annualized loss expectancy (ALE)**
  - **Single loss expectance (SLE) x annualized rate of occurrence (ARO) = ALE**
- **Annualized rate of occurrence (ARO)**
  - **On an annualized basis, the frequency with which a threat is expected to occur**
- **Exposure factor (Impact = asset value x exposure factor)**
  - **A measure of the magnitude of loss or impact on the value of an asset**
- **Probability**
  - **Chance or likelihood, in a finite sample, that an event will occur or that a specific loss value may be attained should the event occur**

Cost/Benefit Analysis

Expected Loss = probability of loss x amount of loss
= probability of loss x impact

Annual Expected Loss = Annual Rate of Occurrence x Impact
= ARO x Asset value x Exposure Factor

If the cost of defending your assets is less than the expected loss if you don't defend them, you should invest in security

More on Expected Loss

- Suppose we have a set of valuable information assets $\{\alpha_k\}$
- Suppose there are a set of threats $\{T_j\}$
- Let v_k be the value of α_k and e_{jk} be the exposure factor for asset α_k when α_k is successfully attacked by T_j
- Let p_{jk} be the probability of a successful attack on α_k by T_j
- Then the single loss expectance due to a successful attack is given by

$$SLE = (v_k \times e_{jk}) \times p_{jk}$$

More on Probability of Loss

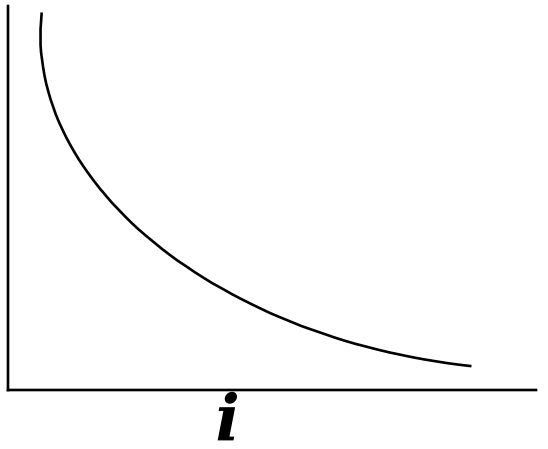
- Let p_{jk} be the probability of a successful attack given the current state of our security, and p_{jk}^{i} be the probability of a successful attack if we make an investment *i* that enhances our security
- If $p_{jk} = 0$, the asset is invulnerable and no investment is needed
- If $p_{jk} = 1$, the asset is completely exposed and investments may be pointless
  - (When is this assumption valid? When is it not valid?)
- If $0 < p_{jk} < 1$, careful investment may reduce vulnerabilities, so $p_{jk}^{i} < p_{jk}$,which, in turn, reduces expected losses
- Since no amount of investment can make an insecure asset completely secure, if $0 < p_{jk} < 1$, then $0 < p_{jk}^{i} < 1$

More on Probability of Loss

- **In general, we expect that increased investments in security would enhance security, although perhaps at a decreasing rate**

- **Thus, $p_{jk}^{i} \rightarrow 0$ as $i \rightarrow \infty$**

- **$\frac{d\ p_{jk}^{i}}{di} < 0$ and $\frac{d^2\ p_{jk}^{i}}{di^2} > 0$**



Daniel J. Ryan
ryand@ndu.edu

Benefits of Investing in Security

- The expected benefit of an investment i in security is

  Expected benefit of i = $(v_k \times e_{jk}) \times [p_{jk} - p_{jk}^i]$

- Deducting the cost of the investment gives us the expected net benefit

  Expected net benefit of i = $(v_k \times e_{jk}) \times [p_{jk} - p_{jk}^i] - i$

- Where the difference between benefits and costs are maximized, the investment is optimal
  - That is, when the expected net benefit is maximized, i is optimal
  - (Does an optimal investment always exist?)

Total Expected Loss

- For an information infrastructure, the expected loss is a summation over all willing and capable threats and all information assets of the product (v_k x e_{jk}) x p_{jk}

$$\text{Expected loss} = \sum_T \sum_\alpha (v_k \times e_{jk}) \times p_{jk}$$

- This comports well with the risk management equation, since if there is no threat, our expected loss disappears
- Thus, p_{jk} is directly proportional to the nature and scope of our vulnerabilities and inversely proportional to the extent to which we effectively employ countermeasures

Now the Bad News

- **We have no scientifically valid statistics upon which to base estimates of the p_{jk}**
- **So we cannot calculate expected losses**
- **So we cannot do quantitative risk management today**
- **So, how do we get the probability distributions we need in order to make decisions based on quantitative risk management?**

Analysis of Failure Time Data

Daniel J. Ryan
ryand@ndu.edu

National Defense University

Resources

- **NIST Engineering Statistics Handbook**

http://www.itl.nist.gov/div898/handbook/index.htm

- **Introduction to Life Data Analysis**

http://www.weibull.com/LifeDataWeb/lifedataweb.htm

Functions

- The *Survivor function* S(t) tells us the probability of being operational at time t.
  - $S(t) = Pr[T \geq t] = Pr[T > t]$

- The *Failure function* tells us the probability of having failed at time t.
  - $F(t) = Pr[T < t] = Pr[T \leq t]$
  - So, $F(t) = 1 - S(t)$
  - Remember $Pr[T = t] = 0$

- The *failure density function* $f(t) = F(t + \delta t) - F(t)$.

Calculations

- Let N be the sample size (e.g. N=1,023,102)
- n(t) = number of system failures prior to age t
- n(t + δ) = number of system failures prior to age t + δ

- [n(t + δ) – n(t)] / N is the portion of the sample that is expected to fail during the interval [t , t + δ) and is equal to F(t + δ) – F(t).

- Thus, $f(t) \approx [n(t + \delta) - n(t)] / \delta \bullet N$

Failure Rate

- The *failure rate function* r(t) is the probability of death per unit time at age t for an individual alive at time t. For small δ, the quantity r(t) • δ is given by

$$r(t) \bullet \delta = \frac{\text{number of deaths during } [\, t \,,\, t+\delta \,)}{\text{number surviving at age } t}$$

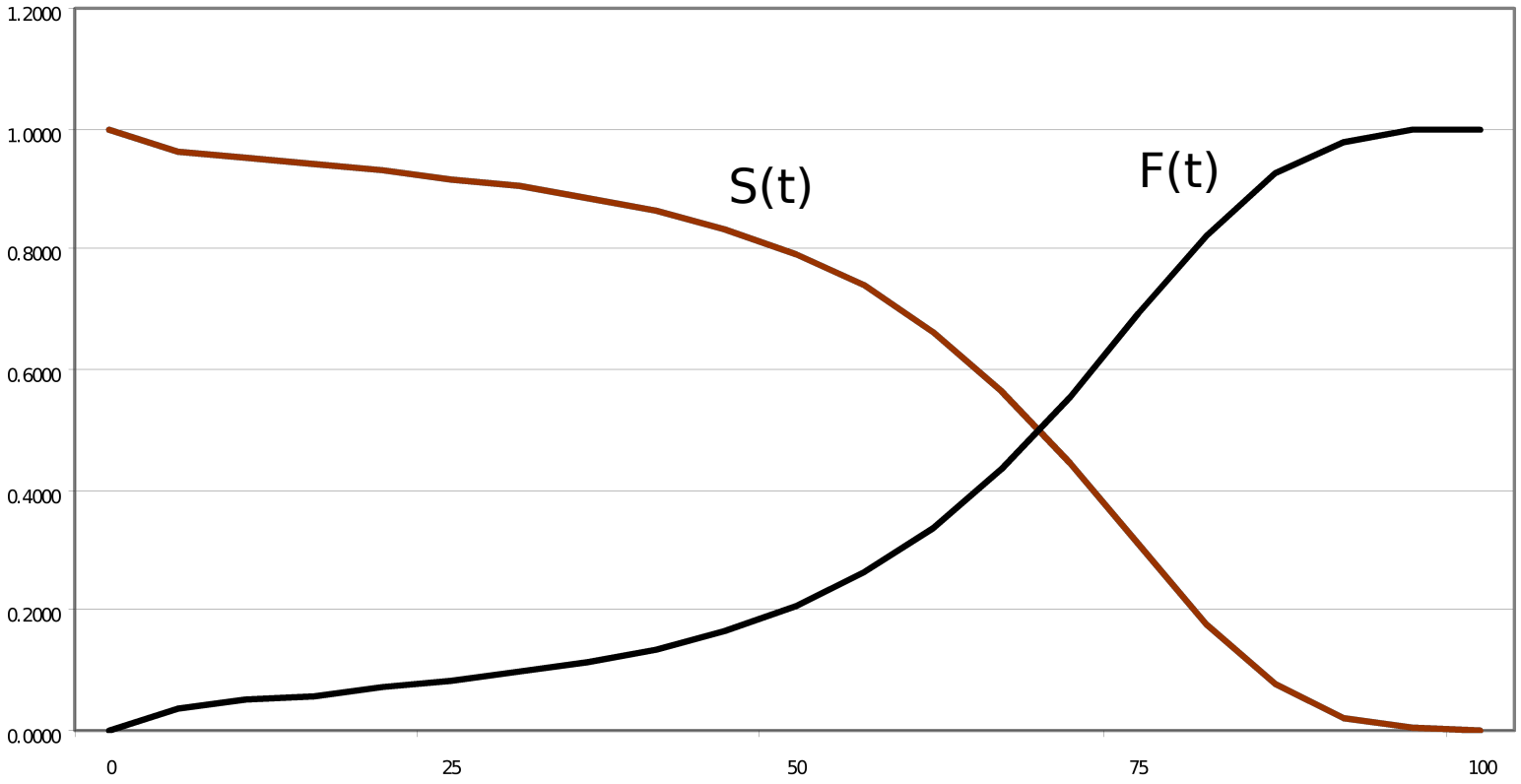
$$= [n(\, t+\delta \,) - N(t)] \,/\, L(t)$$

- Dividing top and bottom by N, we get

$$r(t) \bullet \delta = [\, f(t) \bullet \delta \,] \,/\, S(t) \text{ so } r(t) = f(t)/S(t)$$

- The failure rate is also called the *hazard function*

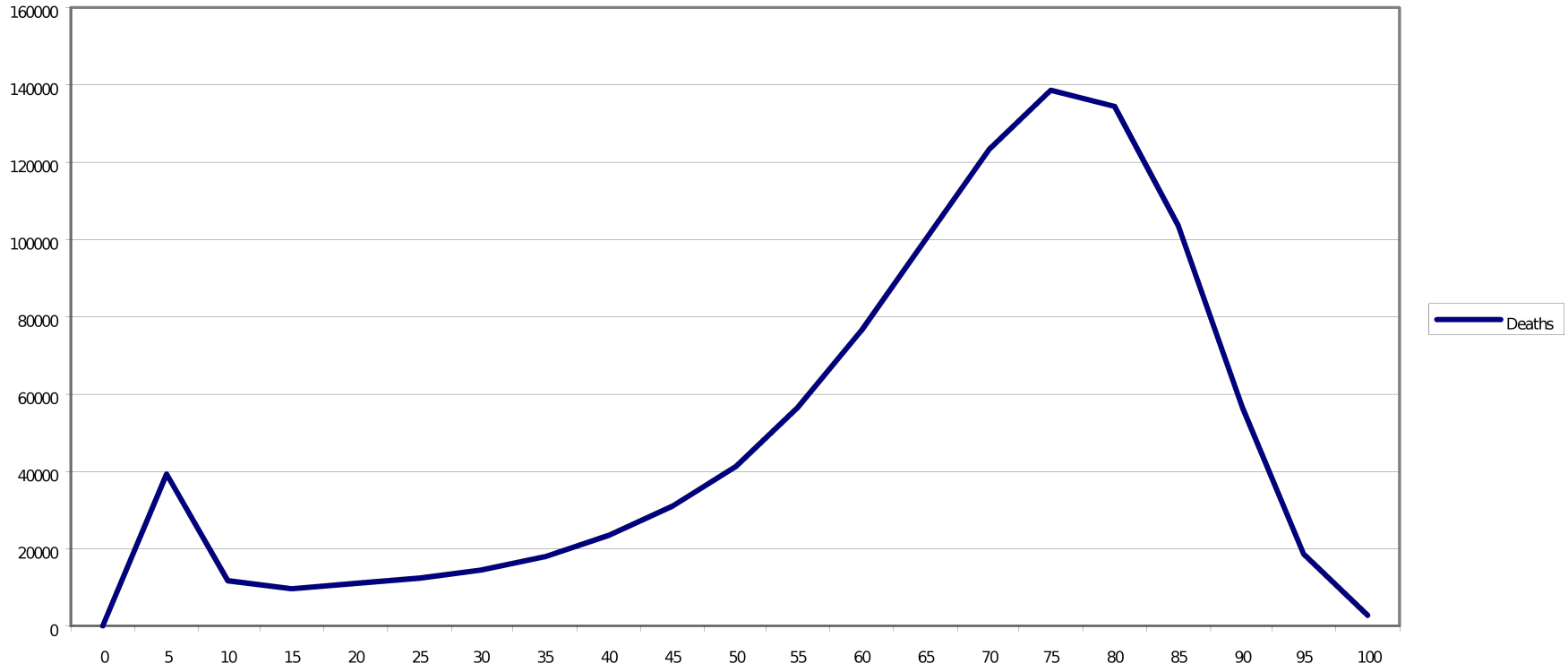
Survival and Failure Distributions



Daniel J. Ryan
ryand@ndu.edu

National Defense University

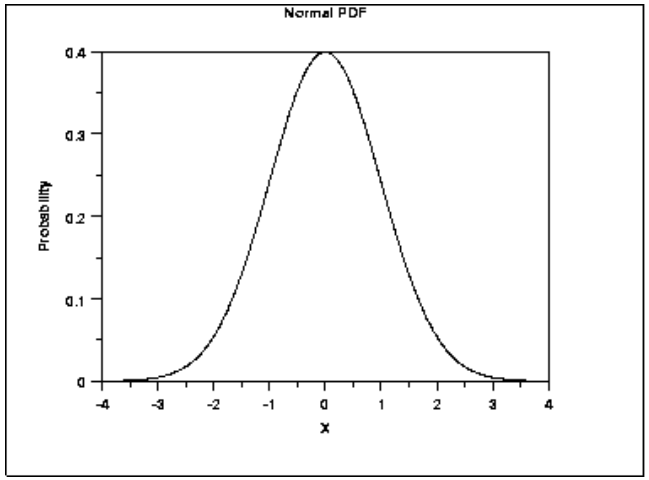
Failure Rate

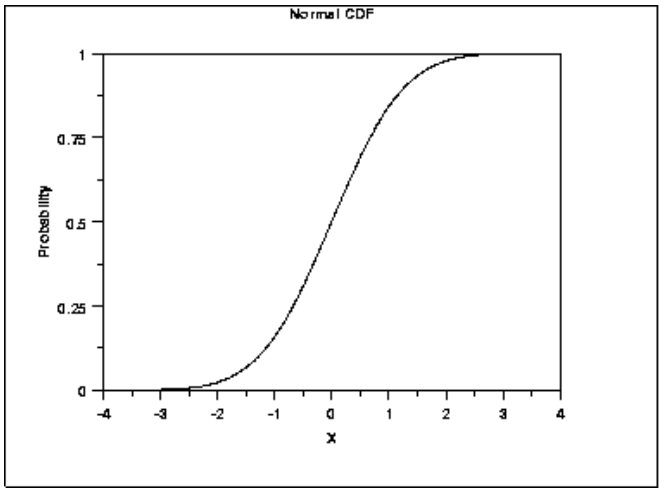


Daniel J. Ryan
ryand@ndu.edu

National Defense University

Normal Distribution





The normal distribution is widely used. It is well-behaved and mathematically tractable.

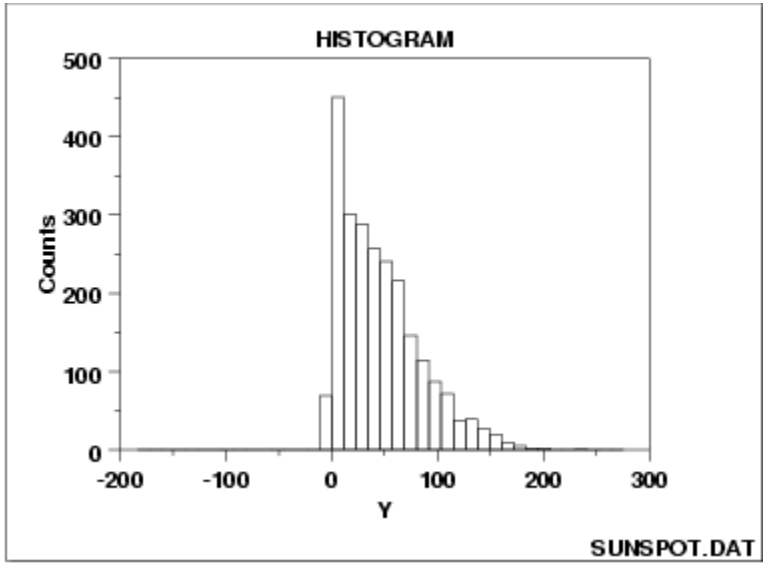
The central limit theorem says that as the sample size gets large (1) the sampling distribution of the mean is approximately normal regardless of the distribution of the original variable, and (2) the sampling distribution of the mean is centered at the mean of the original variable, while the standard deviation of the sampling distribution of the mean approaches $\overline{s}$ $\sigma/\sqrt{N}$

http://www.itl.nist.gov/div898/handbook/eda/section3/eda36

Daniel J. Ryan
ryand@ndu.edu
National Defense University

Skew

- **Unfortunately, in analysis of failure times, the data are not usually normally distributed. The bulk of the data appear at the left end of the distribution and the distribution has a longer tail to the right. We say that such distributions are *right skewed*. In skewed distributions, the median is significantly displaced from the mean.**

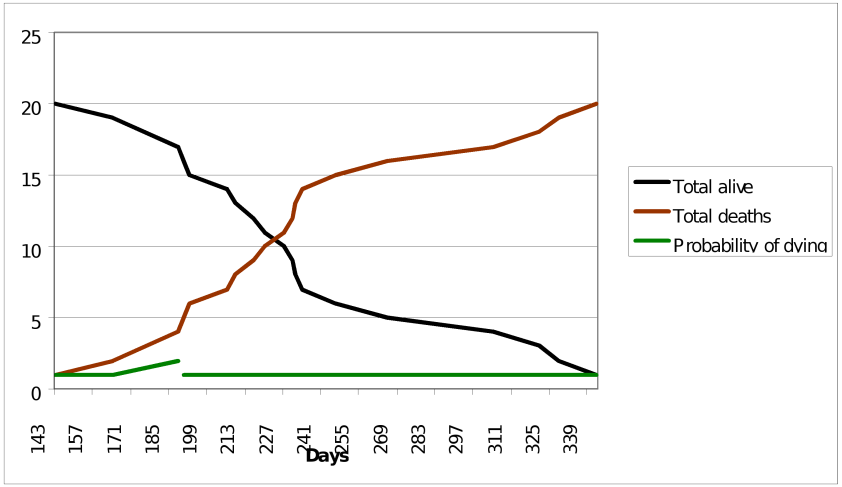


http://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm

Creating Distributions from Survival Data

143	164	188	188	190
192	206	209	231	216
220	227	230	234	246
265	304	321	328	342

N = 20



Daniel J. Ryan
ryand@ndu.edu

National Defense University

Censoring

N = 20

143	**164**	**188**	**188**	**190**
192	**206**	**209**	**231**	**216**
220	**227**	**230**	**234**	**246**
265	**304**	**321 (216)**	**328 (244)**	**342 (325)**

- **When some individuals do not fail during the observation period (are alive at the end of the study), they are said to be *right censored*.**
- **In this example, the last three systems, which would have died on days 321, 328 and 342, instead disappeared from the study on days 216, 244, and 325, respectively. These three are right censored.**

Censoring (continued)

- **When censored survival times are observed only if failure had not occurred prior to a predetermined time at which the study was to be terminated, or the individual has a specific fixed censoring time, the censoring is *Type I censoring*.**

- ***Type II censoring*, or *statistic censoring*, occurs where the study terminates as soon as certain order statistics are observed (e.g. a specified number of failures have occurred).**

- **Sometimes, inferential procedures are easier for type II than for type I censoring, but**
  - **Type II does not allow an upper bound on study duration**
  - **Cannot be used if there is staggered entry into the study**

Kaplan-Meier Estimator

- **The function**

  $$F_n(t) = [\text{ number of sample values} \leq t \text{ }] / n$$

  Is an estimator for the distribution function $F(t) = Pr[\ T \leq t\ ]$

  It is called the empirical estimation function.

- **If there are no censored values, the sample survivor function $S_n(t) = 1 - F_n(t)$ is a step function that decreases by 1/n at each failure time observed.**

- **Unfortunately, this doesn't work if there are censored individuals in the sample, so another method is needed.**
- **A generalization of the sample survivor function for censored data was presented by Kaplan and Meier in 1958.**

Kaplan-Meier Estimator

- Suppose $t_1 < t_2 < \ldots < t_k$ are observed failure times from a homogenous group with an unknown survivor function F(t).
- Suppose d_j individuals fail at time t_j and that c_j individuals are censored in the interval $[t_j, t_{j+1})$ at times $t_{j1}, t_{j2}, \ldots, t_{jm}$, for j=0, 1, 2, …, k.
- Let $t_0 = 0$ and $t_{k+1} = \infty$.
- Let n_j be the number of individuals at risk at a time t_j^-just prior to t_j.
- The probability of a failure at t_j is

$$\Pr[\, T = t_j \,] = F(t_j^-) - F(t_j)$$

Kaplan-Meier Estimator

- Assume that the contribution to the liklihood of a censored survival time at t_{jl} is $Pr[T > t_{jl}] = F(t_{jl})$
- We are assuming that the censoring mechanism is independent.
- The Kaplan-Meier estimate, or product limit estimate, of the survivor function is

$$S^{KM}(t) = \prod_{j|tj<t} (n_j - d_j)/n_j$$

- The KM estimator makes the estimated hazard, or conditional probability of failure, at each t_j agree exactly with the observed proportion d_j/n_j of the n_j individuals at risk that fail at time t_j.
- $S^{KM}(t)$ is undefined for times greater than t_{kmk}.

Hazard Functions

- The *hazard function* h(t) is the probability that an individual fails at time t, given that the individual has survived to that time.

$$h(t) = \lim_{\delta t \to 0} \Pr[\, t \le T < t+\delta t \mid T \ge t \,] / \delta t \,]$$

- h(t) δt is the approximate probability that an individual will die in the interval (t, t+δt), having survived up until t.
- The hazard function is usually interpreted as the risk of failure at time t.

- h(t) = f(t)/S(t) and S(t) = exp{ -H(t)}, where H(t) = $\int_0^t h(u)du$ is the *cumulative hazard function*

The Nelson-Aalen Estimator

- **The Nelson-Aalen estimator of the survivor function S(t) is given by**

$$S^{NA}(t) = \prod_{j=1, 2, \ldots k} \exp(- d_j)/n_j$$

- **The Nelson-Aalen estimator is also known as Altshuler's estimate.**
- **The Kaplan-Meier estimator is an approximation to the Nelson-Aalen estimator.**
- **The Nelson-Aalen estimator will always be greater than the Kaplan-Meier estimator at any time t.**
- **The Nelson-Aalen estimator performs better than the Kaplan-Meier estimator for small samples, but in most cases they are very similar.**

Hazard Functions

- $H(t) = - \sum \log[(n_j - d_j)/n_j]$ for the KM estimator
- $H(t) = \sum d_j/n_j$ for the KA estimator
- Since $h(t) = [H(t+\Delta) - H(t)]/ \Delta$
  solving for h(t) gives

$$h(t) = d_j/(n_j \blacksquare \Delta_j), \text{ where } \Delta_j = t_{j+1} - t_j$$

Daniel J. Ryan
ryand@ndu.edu
National Defense University

The Cox Model

Daniel J. Ryan
ryand@ndu.edu

National Defense University

Non-Parametric Studies

- Non-parametric methods are useful in
  - Analysis of a single sample of data
  - Comparisons of two or more groups of survival times
- Not useful when the population is not homogenous
  - Some computers are behind firewalls; some not
  - Some systems have AV; some don't
  - Some systems are C2; some B1
- Not useful when there are more than one type of hazard
  - Hackers
  - Malicious code
  - Denial of service attacks

Modeling

- In analyses of survival time, the focus is on the time interval during which the systems being studied survive, up until the time at which they are successfully attacked
- In modeling survival data, we focus on the hazard function
- Two objectives:
  - Determine which explanatory variables affect the form of the hazard function, and to what extent
  - To estimate the hazard function for an individual system

Proportional Hazards

- **Proposed by Cox in 1972.**
- **The Cox model is a *semi-parametric model*, since no specific distribution is assumed for the survival times**
- **Let h (t; *i*) describe the hazard function for our current IT infrastructure following an investment *i* in information security**
- **Suppose**

  h (t; *i*) = c · h (t; 0), where c is a constant and 0 < c < ∞
- **The post-investment hazard is proportional to the current hazard**

Proportional Hazards (cont.)

- **The constant of proportionality, c, is the ratio of the hazards of death for an individual system that has the "benefit" of the investment to the hazard facing an individual who does not have that "benefit"**

- **c is called the *hazard ratio***
  - **If c < 1, the hazard is less for an individual protected by the investment**
  - **If c > 1, the hazard has actually increased for the individual following the investment**

Proportional Hazards (cont.)

- Suppose we have k systems in our infrastructure.
- Let $h_j(t ; i)$ be the hazard function for the j^{th} system,
  j = 1, 2, …, k, following an investment *i* in information security
- Then $h_j(t ; i) = c \cdot h_j(t ; 0)$
- Since 0 < c, let $\beta = \log c$
- Any value of $\beta = (-\infty, \infty)$ will yield a positive c
- Positive values of β occur when c > 1, and the investment is detrimental, rather than desirable

Explanatory Variables

- Let X be an indicator variable that indicates whether an individual system is protected by the investment *i*
- Let x_j be the value of X for the j^{th} system in the study
- Let x_j be 1 if system *j* is protected; zero otherwise
- Then

  $h_j(t ; i) = e^{\beta x_j} \cdot h_j(t ; 0)$

- This is the proportional hazards model for comparing two groups of systems, one which consists of systems protected by *i*; the other not

Explanatory Variables (cont.)

- This model can be generalized by using a vector of explanatory variables in place of a single indicator variable
- Suppose the hazard of failure at a particular time depends on the values $x_1, x_2, \ldots, x_p$ of explanatory variables
  $X_1, X_2, \ldots, X_p$
- Let a vector $\chi_j = (x_{1j}, x_{2j}, \ldots, x_{pj})$ describe the status of system *j* with respect to the p explanatory variables
- Then $h_j(t ; i) = c(\chi_j) \cdot h_j(t ; 0)$, where $c(\chi_j)$ is a vector function of the values of χ_j

Explanatory Variables (cont.)

- Since $0 < c(\chi_j)$, we can write $c(\chi_j) = \exp(\eta_j)$, where $\eta_j = \beta_1 \cdot x_{1j} + \beta 2 \cdot x_{2j} + \ldots + \beta_p \cdot x_{pj}$
- In matrix notation
  $\eta_j = \beta' \cdot x_j$, where β is the vector of coefficients of the explanatory variables $x_1, x_2, \ldots, x_p$
- The general proportional hazards model then is

$$h_j(t ; i) = \exp(\beta_1 \cdot x_{1j} + \beta 2 \cdot x_{2j} + \ldots + \beta_p \cdot x_{pj}) \cdot h_j(t ; 0)$$

- There are other choices for $c(\chi_j)$, but $c(\chi_j) = \exp(\beta' \cdot x_j)$ is among the best known and most widely used

Conclusions

- By drawing on models from the medical community's approach to measuring the value of proposed drugs and drug protocols, we can estimate the probability distributions needed to calculate expected losses
- By using a proportional hazards approach, we can evaluate the contribution to security of a variety of design and operational factors
- But . . . Failure time data needs to be collected using double-blind studies to drive the mathematical models we now have